Data Interoperability at the IRI: translating between data cultures

Benno Blumenthal

International Research Institute for Climate Prediction

Columbia University

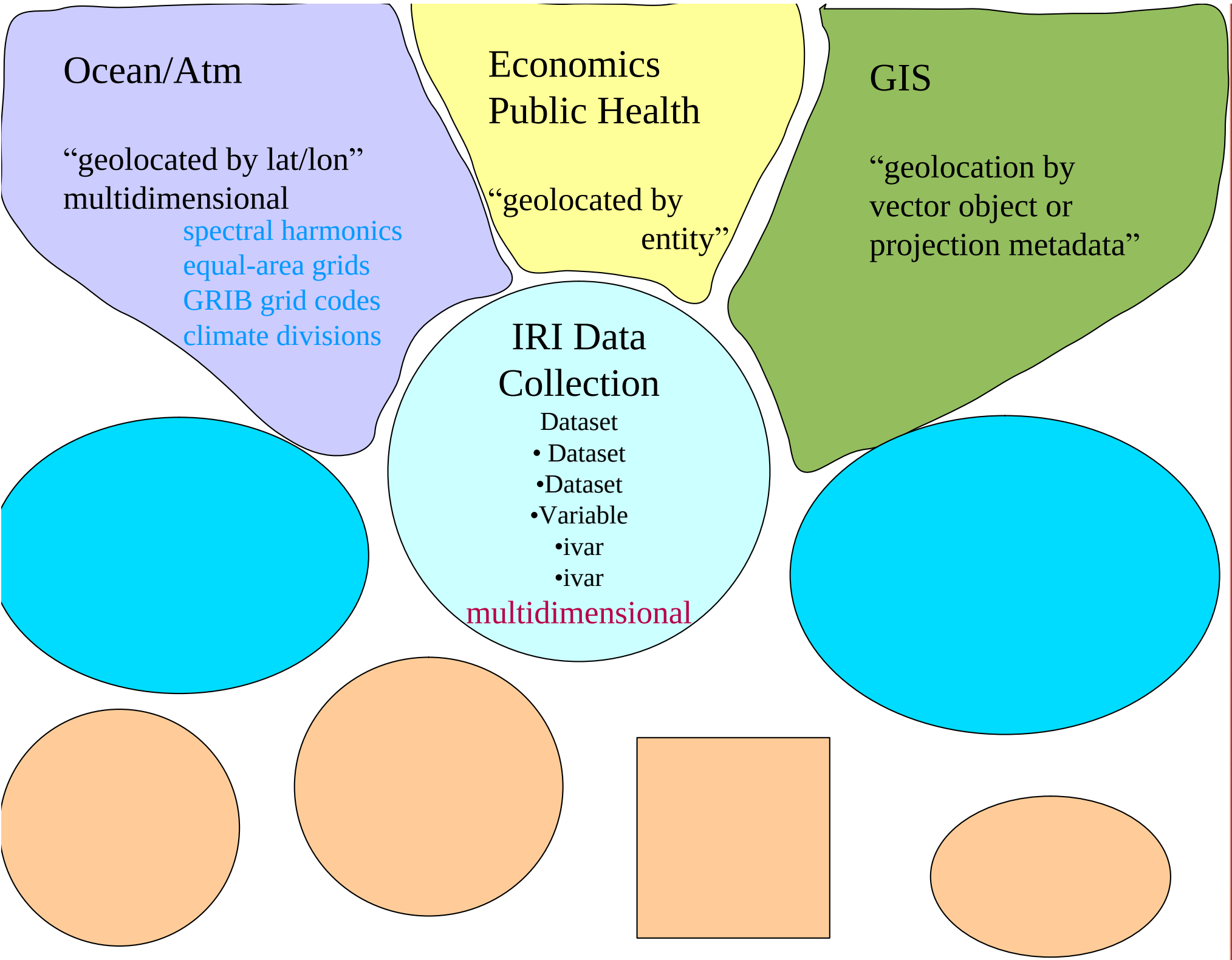
INTERNATIONAL RESEARCH INSTITUTE

Work presented has been done by

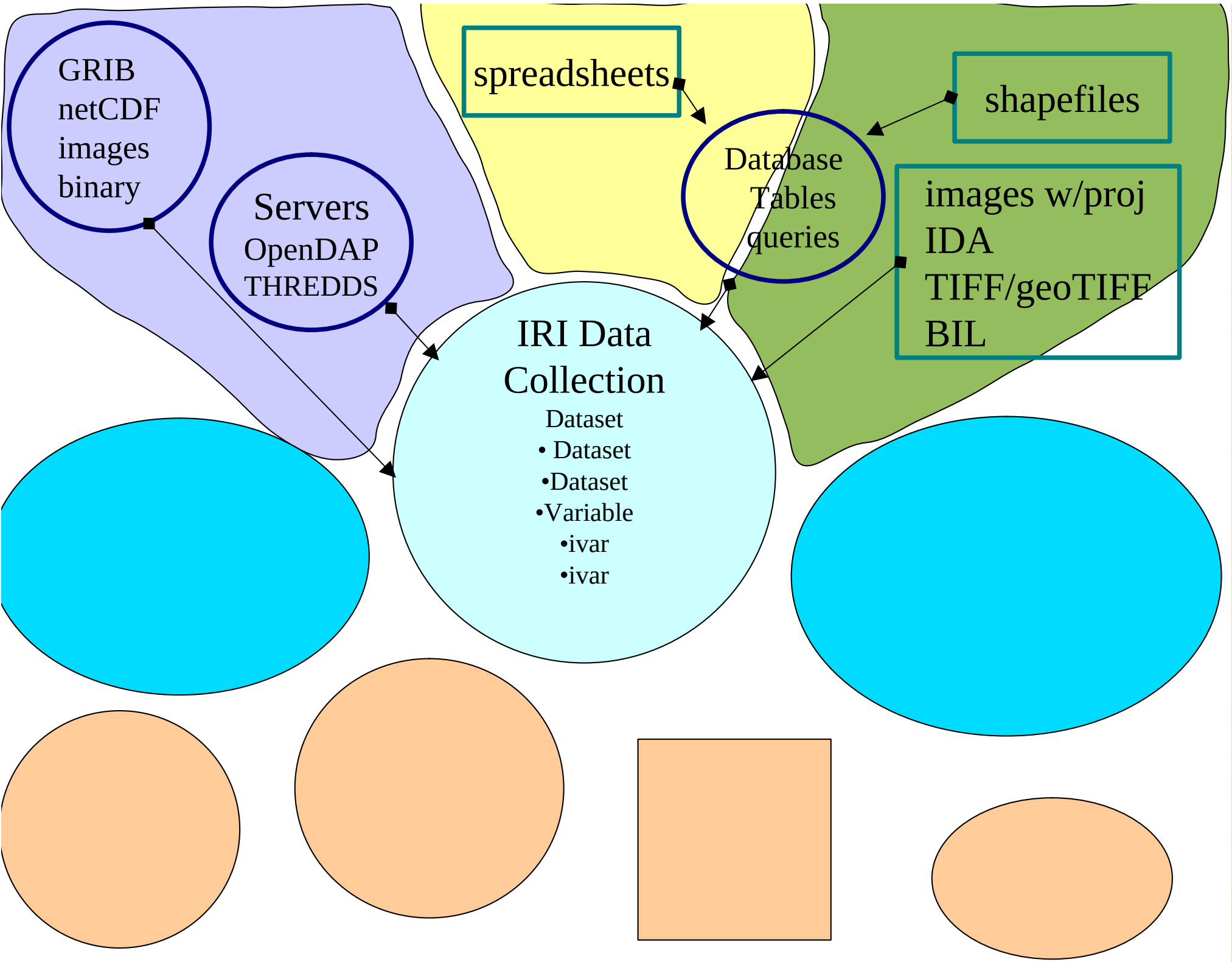
- John del Corral
- Michael Bell
- Emily Grover-Kopec

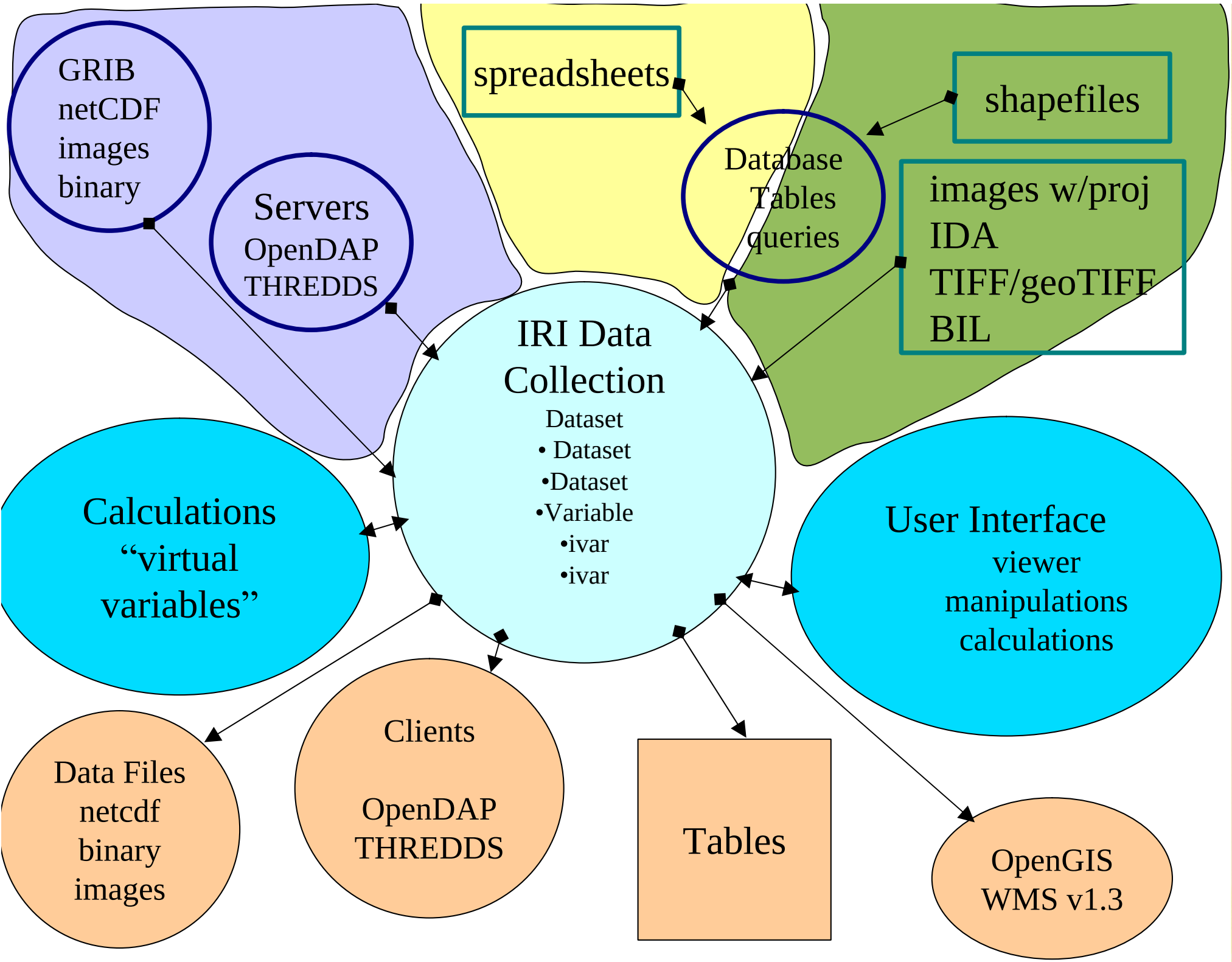
Outline

- Data Cultures
- Needed Functionality for usefully merging
- Data model changes
- Functionality changes
- Example
- Dataset Connections
- Mapping SimpleFeatures to OpenDAP



INTERNATIONAL RESEARCH INSTITUTE




GRIB
netCDF
images
binary
Servers
OpenDAP
THREDDS
spreadsheets
Database
Tables
queries
shapefiles
images w/proj
IDA
TIFF/geoTIFF
BIL
IRI Data
Collection
Dataset
• Dataset
•Dataset
•Variable
•ivar
•ivar
Calculations
“virtual
variables”
User Interface
viewer
manipulations
calculations
Data Files
netcdf
binary
images
Clients
OpenDAP
THREDDS
Tables
OpenGIS
WMS v1.3
INTERNATIONAL RESEARCH INSTITUTE

Functionality for new Users

- want to use our current data holdings
- aggregation (particularly of images)
- time analysis of GIS data
- translation to “entity” basis

Data Model Changes

- projection attributes
  - SpatialReferenceSystemWKT (An OpenGIS standard)
  - SpatialReferenceSystemDims
- geometry data object
  - OpenGIS Simple Feature

Functionality Changes

- geometry display: fill, fillby, stroke, mark, ...
- projection functionality: objects with differing projection attributes can be made to match
- rasterization: geometry object converted to raster
  - computes the fraction of each gridbox that is covered by the given geometry object

Example

- NDVI[x y time]

  Albers projection

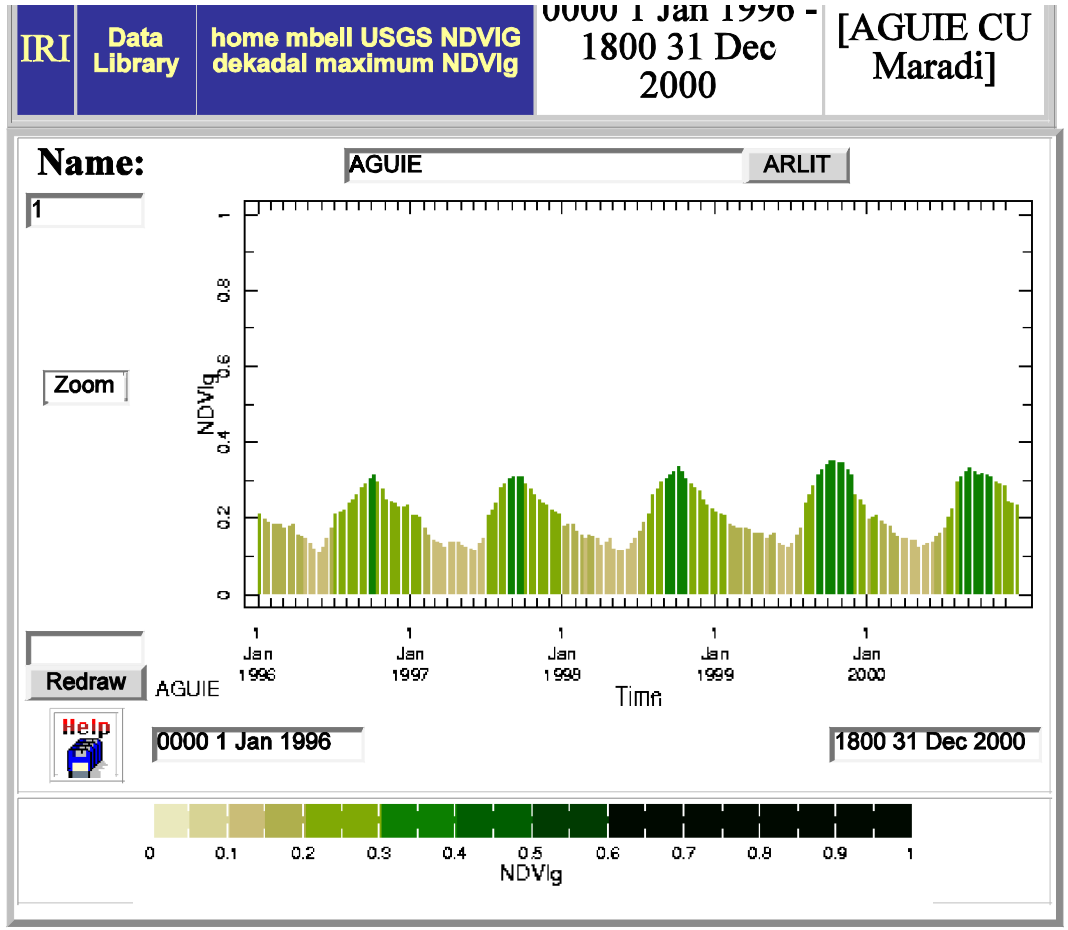
  and

location [d_name]

  Lon/lat geometry

  produces

NDVIg[d_name time]



Get Data	Entire Dataset	data in view	Edit	plot
Page Formats	documented page	plain page	cut and paste link	more options

weighted-average

weighted-average

Computes weighted average

var {Wghts} [grids ...] minfrac **weighted-average**

Arguments		
var	variable	variable to be averaged
Wghts	variable or proc	weights to be used in averaging. If a geometry, will be rasterized to match **var**. If a lat/lon geometry, then a cosine latitude weighing is also applied.
grids	grid set	grids to be averaged over
minfrac	number	Minimum fraction of data that must be present (i.e., fraction not indicated as missing) within the selected domain in order for the weighted-average to be calculated. If *minfrac* is not present, then a missing value is returned. If *minfrac* is not given, then the weighted-average is calculated regardless of the amount of data present in domain. (optional)
Returns		
wtavgvar	variable	*var* averaged over *grids* weighted by *Wghts*. *wtavgvar* no longer depends on *grids*, but it does depend on any other grids that *var* and/or *Wghts* depend on.

Example: NINO3.4 region

SOURCES .NOAA .NCDC .ERSST .version2 .SST 90 -5 240 5 georect [X Y]weighted-average

Dataset connections

Metadata standards are useful because they let us make meaningful connections between datasets.

Connections we might want

NDVI(district,time)

Difficult to document dataset – currently at IRI user has to follow a series of links to find out about districts and related information

Would be nice if

- User searches
- Software searches, e.g. data viewer could display district information, locator map …
- Lists of related information for user's next step

Two kinds of Dataset Connections

1. Standard variables
2. Standard units
3. Definitional connections (e.g. same dataset or table)
4. Computational connections (data connected by construction)

Latter are local but solid; former are standard and possibly hazy

Metadata Interoperability

- Attributes (all agree)
- Basic XML (all agree)
- XML with name spaces

Given a dataset following one metadata standard, how do we make connections to datasets following a different standard?

Conceptual Model

- Standard concepts and standard relationships
- Delayed mapping between standards means the information loss in impedance mismatches does not propagate

Semantic Web

- XML (with namespaces) readable files with tagged data
- XML Schema structure of a tagged file
- RDF standard for expressing relations
- RDF Schema standard for expressing relations between classes
- OWL **logical** derivations of relations – our relations are different, e.g. projection transformation to connect data in different projections

Sample Geometry as OpenDAP

Example: climate division outlines associated with climate division data are conceptually

Location[IDIV]

Where location is a geometry (multipolygon)

This becomes nested sequences in OpenDAP

- dds
- das

with some attributes

Simple Features and OpenDAP v2.0 structure

Point	structure {float lat;float lon;} point;
LineString	sequence {float lat;float lon;}LineString;
MultiPoint	sequence {float lat;float lon;} MultiPoint;
Polygon	sequence {sequence {float lat;float lon;} aring;} Polygon;
MultiLineString	sequence {sequence {float lat;float lon;} LineString; }MultiLineString;
MultiPolygon	sequence {sequence {sequence {float lat;float lon; } ring;} Polygon;} MultiPolygon;
Geometry Collection	sequence {string OpenGISSimpleFeature; sequence{...}geom }GeometryCollection; but how to handle a collection of a collecton???

Simple Features and OpenDAP v2.0 attributes

- SpatialReferenceSystemWKT

  Projection information

- OpenGISSimpleFeature

  Point, LineString, Polygon, MultiPoint, MultiLineString,MultiPolygon,GeometryCollection

Or

- Dimensionality
  - *0 (point) 1 (line) 2 (polygon)*

Sample Attributes

NDVI [x y time]

SpatialReferenceSystemWKT

PROJCS["Albers_Equal_Area_Conic",GEOGCS["GCS_North_American_1927",DATUM["D_North_American_1927",SPHEROID["Clarke_1866",6378206.4,294.9786982]],PRIMEM["Greenwich",0],UNIT["Degree",0.0174532925199433295]],PROJECTION["Albers"],PARAMETER["False_Easting",0],PARAMETER["False_Northing",0],PARAMETER["Central_Meridian",20],PARAMETER["Standard_Parallel_1",21],PARAMETER["Standard_Parallel_2",-19],PARAMETER["Latitude_Of_Origin",1],UNIT["Meter",1]]

SpatialReferenceSystemDims

x y

Lessons Learned

- Must be able to specify different standards for different parts of a dataset – country codes are an iso standard, for example, but are unlikely to ever be part of CF or WCS
- Must be able to transmit everything, not just anointed datasets
- Bringing together different concepts in different data cultures (multidimensionality, location as object) improves both sides of the divide.